

Prepositional Attachment Disambiguation Using Bilingual Parsing and Alignments

Geetanjali Rakshit^{♣♥} Sagar Sontakke[♣] Pushpak Bhattacharyya[♣] Gholamreza Haffari[♥]
[♣]IITB-Monash Research Academy, IIT Bombay [♥]Monash University
[♣]Dept. of Computer Science and Engineering, IIT Bombay Australia
geet, sagarsb, pb@cse.iitb.ac.in first.last@monash.edu

Abstract

In this paper, we attempt to solve the problem of Prepositional Phrase (PP) attachments in English. The motivation for the work comes from NLP applications like Machine Translation, for which, getting the correct attachment of prepositions is very crucial. The idea is to correct the PP-attachments for a sentence with the help of alignments from parallel data in another language. The novelty of our work lies in the formulation of the problem into a dual decomposition based algorithm that enforces agreement between the parse trees from two languages as a constraint. Experiments were performed on the English-Hindi language pair and the performance improved by 10% over the baseline, where the baseline is the attachment predicted by the MSTParser model trained for English.

1 Introduction

Prepositional Phrase (PP) attachment disambiguation is an important problem in NLP, for it often gives rise to incorrect parse trees. Statistical parsers often predict incorrect attachment for prepositional phrases. For applications like Machine Translation, incorrect PP-attachment leads to serious errors in translation. Several approaches have been proposed to solve this problem. We attempt to tackle this problem for English. English is a syntactically ambiguous language with respect to PP attachments. For example, consider the following sentence where the prepositional phrase *with pockets* may attach either to the verb *washed* or to the noun *jeans*.

Sentence(1): I washed the jeans with pockets.

Below is the correct dependency parse tree (for sentence 1) where the prepositional phrase *with pockets* is attached to the noun *jeans*.

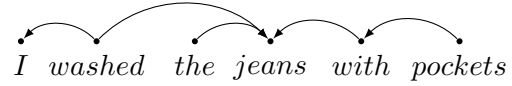


Figure 1: Dependency Parse Tree for Sentence 1

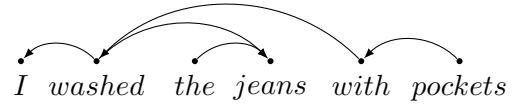


Figure 2: Incorrect Parse Tree for Sentence 1

Another possible parse tree for the same sentence could be as shown in Figure 2: A statistical parser often predicts the PP-attachment incorrectly, and may lead to incorrect parse trees. Let us now look at another sentence.

Sentence(2): I washed the jeans with soap.

The correct dependency tree for sentence [2] is the following (Figure 3), where the prepositional phrase *with soap* attaches to the verb *washed*.

Clearly, there is a case of ambiguity that can be resolved only if the semantics are known. In this case, the fact that *soap* is an aid to the verb *washed* disambiguates its attachment to the verb rather than the noun *jeans*. For correctly translating such an English sentence to another language, the attachments need to be marked correctly.

In this work, we propose a Dual Decomposition (DD) based algorithm for solving the PP attachment problem. We try to disambiguate the PP attachments for English using the corresponding parallel Hindi corpora. Hindi is a syntactically

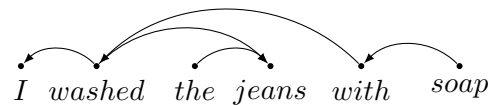


Figure 3: Dependency Parse Tree for Sentence 2

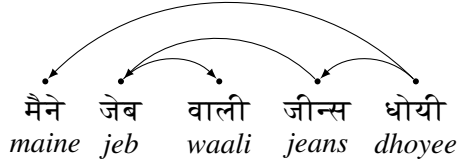


Figure 4: Dependency Parse Tree for Sentence 3

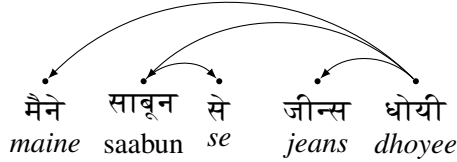


Figure 5: Dependency Parse Tree for Sentence 4

rich language and in most cases exhibits no attachment ambiguities. The use of case markers and the inherent construction of sentences in Hindi make cases of ambiguity rarer. Let us examine how sentences 1 and 2 would look like in Hindi, and if there is a case for ambiguity.

Sentence (3) and sentence (4) are the respective Hindi translations of sentence (1) and (2)).

Sentence (3): मैंने जेब वाली जीन्स धोयी |

Sentence (4): मैंने साबून से जीन्स धोयी |

In sentence (3), the prepositional phrase *jeb waali* attaches to the noun *jeans* as shown in the figure 4.

The parse tree for sentence (4) is shown in figure 5, where the prepositional phrase *saabun se* attaches to the verb *dhoyee*.

The case markers *waali* and *se* in the two sentences in Hindi make the pp-attachment clear. In our approach, we make use of the parallel Hindi sentences to disambiguate the PP attachments for English sentences.

The roadmap of the paper is as follows: We discuss the literature and related work for solving the PP-attachment problem in section [2]. Section [3] describes our approach, and the Dual Decomposition algorithm in detail. The setup, data, and experiments are covered in Section [4]. With Section [5], we conclude our work and discuss scope for future work.

2 Related Work

A number of supervised and unsupervised approaches for solving the PP-attachment problem have been proposed in the literature. Ratnaparkhi

et al. (1994) use a Maximum Entropy Model for solving the PP-attachment decision. Schwartz et al. (2003) propose an unsupervised approach for solving PP attachment using multilingual aligned data. They transform the data into high-level linguistic representations and use it make reattachment decisions. The intuition is similar to our work, but the approach is entirely different. Brill and Resnik (1994) discuss a transformation-based rule derivation method for PP-attachment disambiguation. It is a simple learning algorithm which derives a set of transformation rules from training corpus, which are then used for solving the PP-attachment problem. Stetina and Nagao (1997) make use of the semantic dictionary to solve the problem of disambiguating PP attachments. Their work describes use of word sense disambiguation (WSD) for both supervised and unsupervised techniques. Agirre (2008) and Medimi (2007) have used WSD-based strategies in different capacities to solve the problem of PP-attachment. Olteanu and Moldovan (2005) have attempted to solve the pp-attachment problem as a classification problem of attachment either to the preceding verb or the noun, and have used Support Vector Machines (SVMs) that use complex syntactic and semantic features.

3 Our Approach

We propose a Dual Decomposition based inference algorithm to look at the problem of PP-attachment disambiguation. Dual decomposition, or more generally, Lagrangian Relaxation, is a classical method for combinatorial optimization and has been applied to several inference problems in NLP (Rush and Collins, 2012). We train two separate parser models for English and Hindi each, using the MSTParser, and make use of these models in the inferencing step. The input to the algorithm is a parallel English-Hindi sentence pair, with its word alignments given. We first obtain the predicted parse trees for the English and Hindi sentences from the respective trained parser models as an initialisation step. The DD algorithm then tries to enforce agreement between the two parse trees subject to the given alignments.

Let us take a closer look at what we mean by agreement between the two parse trees. Essentially, if we have two words in the English sentence denoted by i and i' , aligned to words j and j' in the parallel Hindi sentence respectively, we

can expect a dependency edge between i and i' in the English parse tree to correspond to an edge between j and j' in the Hindi parse tree, and vice versa. Also, in order to accommodate structural diversity in languages (Smith and Eisner, 2006), we can expect an edge in the parse tree in one language to correspond to more than one edge, or rather, a path, in the other language parse tree. This has been captured in the examples in figures 6(A) and 6(B). For an edge in the English parse tree, we term the corresponding edge or path in the Hindi parse tree as the *projection* or *projected path* of the English edge on the Hindi parse tree, and similarly there are projected paths from Hindi to English. For matters of simplicity, we ignore the direction of the edges in the parse trees. The dual decomposition inference algorithm tries to bring the parse trees in the two languages through its constraints.

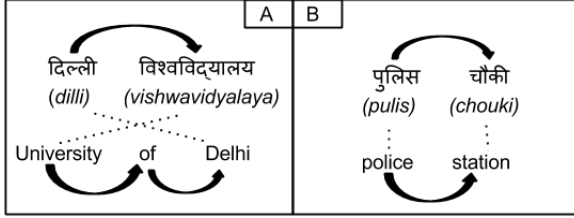


Figure 6: Edge Projection from Hindi to English

The problem is formulated as below:

In the above formulation, e and h represent a English and Hindi sentence respectively. T_e and T_h are the corresponding parse trees. θ_E and θ_H are the model parameters for the edge-factored parser models trained for English and Hindi respectively. t_e represents an edge in the English parse tree T_e . $proj(t_e, T_h)$ is a projected path in Hindi parse tree (T_h) for a given English edge t_e . The term $scr(proj(t_e, T_h))$ stands for the score of a projected path in Hindi parse tree (T_h) for a

$$\begin{aligned} & \underset{t_e, t_h}{\operatorname{argmax}} \left(\log P_{\theta_E}(T_e|e) + \log P_{\theta_H}(T_h|h) + \right. \\ & \left. \sum_{t_e \in T_e} scr(proj(t_e, T_h)) + \sum_{t_h \in T_h} scr(proj(t_h, T_e)) \right) \end{aligned}$$

Figure 7: Optimization Problem

given English edge t_e .

The score of the projected path is calculated as the sum of scores of all edges in the path. Let π_{t_e} denote the projected path on sentence h in Hindi for the edge t_e in the English parse tree. We assume $scr(\pi_{t_e}) = \sum_{a \in \pi_{t_e}} \psi_{t_e}(a)$ where $\psi_{t_e}(a)$ is the score of edge a in the projected path π_{t_e} . In the other direction, π_{t_f} and $scr(\pi_{t_f})$ is similarly defined.

To solve this maximization problem in figure 7, we assume one tree to be given and maximize the other and the score of its projected path. The algorithm is described in detail in section 3.1.

3.1 Dual Decomposition based Algorithm

We use an iterative *Coordinate Descent* algorithm (Algorithm 1) which calls the *Project Algorithm* until convergence. The trees T_e and T_h are initialized by the previously trained parser models for the respective languages.

Algorithm 1 Coordinate Descent Algorithm

- 1: Initialize T_e and T_h from the MSTParser models
 - 2: **for** $t = 1$ to N
 - 3: $T_e^+ \leftarrow project(T_h, e)$
 - 4: $T_h^+ \leftarrow project(T_e, h)$
 - 5: **if** $(T_e == T_e^+) \text{ or } (T_h == T_h^+)$
 - 6: **break**
 - 7: **else**
 - 8: $T_e = T_e^+$
 - 9: $T_h = T_h^+$
 - 10: **end for**
-

For N iterations, the *project* function returns a parse tree for English which maximizes the agreement between the English and Hindi parse tree when the Hindi parse tree is fixed, and likewise for the Hindi parse tree. The algorithm converges when the trees no longer change,

Let us now look at the *Project* algorithm (Algorithm 2) in detail. It predicts the tree for a sentence in the target language, given the parse tree in the source language, and the word alignments between the parallel sentence.

The lagrangian multipliers are initialized to zero. The best tree in the target language is predicted by the argmax computation in step 4. This maximization involves the parser model parameters $\theta(i, j)$ and the score of the best projected path in the source tree for all edges. $r(i, j)$ denotes the score of the projected path of the edge $y(i, j)$ on

Algorithm 2 Project Algorithm (tree T, sen S)

Require: A parse tree T (Hindi) and sentence S (English)

```
1: Initialize  $\forall t, i, j \ u_t(i, j) = 0$ 
2:
3: for  $t = 1$  to  $N$ 
4:    $Y \leftarrow \underset{y \in \text{Trees}(S)}{\text{argmax}} \left( \sum_{i,j} y(i, j) \cdot [\theta(i, j) + \right.$ 
       $\left. r(i, j) - \sum_t u_t(i, j)] \right)$ 
5:
6:   for  $t \in T$ 
7:      $\pi_t \leftarrow \underset{\pi \in \text{paths}(\text{project } t \text{ onto } S)}{\text{argmax}} \sum_{i,j} \pi(i, j) [\varphi_t(i, j, S) + u_t(i, j)]$ 
8:   end for
9:
10:  if  $\forall_{t,i,j}; \pi_t = y(i, j)$  then
11:    return  $Y$ 
12:  else
13:     $\forall_{t,i,j} \ u_t(i, j) \leftarrow u_t(i, j) - \alpha(\pi_t(i, j) - y(i, j))$ 
14:  end for
```

the source tree T. In steps 6 and 7, the best projected path for every edge of the source tree is predicted on the target tree using the classifiers described in section ???. The constraints here are that the edges in the projected paths from the classifiers and the predicted trees are in agreement.

3.2 Projected Path Prediction

In order to predict the projected path in one language for an edge in the other language, we use a set of two classifiers in a pipeline. Let us recall that we have two nodes in one language with an edge between them, and we are trying to predict the path of the corresponding aligned nodes in the other language. The first classifier predicts the length of the projected path, and the second predicts the predicted path itself, given the path length from the first classifier. Let us look at these classifiers separately.

The classifier for path length prediction is a set of five binary classifiers, which predict the path length to be 1, 2, 3, 4 or 5. We assume projected path lengths to be no greater than 5. These classifiers are *perceptrons* trained on separate annotated data. The features used were the words and POS tags of the four nodes in the pair of alignments under consideration.

The classifier for path prediction is a set of four *structured perceptron* classifiers. We train four

classifiers to predict the paths of length 2, 3, 4 and 5. These set of classifiers were trained on separate annotated data, and the features used were the same as in the set of classifiers for path length prediction.

4 Experiments and Results

A parser model was trained for Hindi using the MSTParser (McDonald et al., 2006) by a part of the the Hindi Dependency Treebank data (18708 sentences) from IIIT-Hyderabad (Bhatt et al., 2009). A part of the Penn Treebank (28188 sentences) was used for training an English parser (?). The treebanks were converted to MSTParser format from ConLL format for training. A part of the ILCI English-Hindi Tourism parallel corpus (1500 sentences) was used for training the classifiers. This corpus was POS-tagged using the Stanford POS Tagger (Toutanova et al., 2003) for English and using the Hindi POS Tagger (Reddy and Sharoff, 2011) from IIIT-Hyderabad for Hindi. It was then automatically annotated with dependency parse trees by the parsers we had trained before English and Hindi.

For testing, we created a corpus of 100 parallel sentences and their word alignments from the Hindi-English Tourism parallel corpus. We manually annotated the instances of pp-attachment ambiguity. We examine the prediction for attachment of only these cases. The baseline system used is the attachment predicted by the parser models trained using the MSTParser. We ran experiments on the test set for iterations 10 to 60, in steps of 10. The outputs from the MSTParser trained model and the DD algorithm were compared against the gold data for English.

Our observations have been tabulated in Table 4. The MSTParser model was able to correctly disambiguate 54 number of PP-attachments. Our algorithm, however, performed better and marked 64 number of attachments correctly, in the best case. The baseline accuracy for PP attachment was 54%. With our approach, we were able to achieve an improvement of 10% over the baseline.

We also experimented with the number of iterations to see if the attachment predictions got any better. The observations have been plotted in the graph in figure 8 . Our algorithm performed best at 30 iterations.

In the event of lack of gold standard data for our experiments, we have used statistical POS taggers

Parameter	MST Parser	DD Algorithm
Total Number of PP-attachments	100	100
Number of Correctly identified PP-attachments	54	64
Accuracy (%)	54	64

Table 1: Test Results for English-Hindi

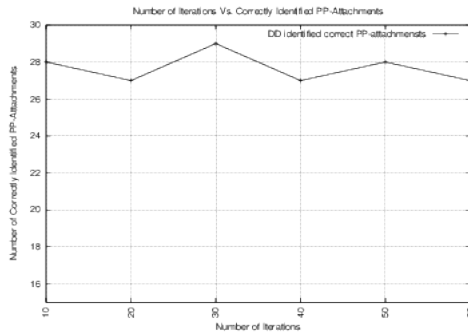


Figure 8: Iterations Vs. Correct PP-attachments

for POS tagging the data. Also, for getting word alignments, we have used GIZA++ (Och and Ney, 2003), which again has scope for errors. These kind of errors may cascade and cause our system to underperform.

5 Conclusion and Future Work

We were able to achieve an accuracy of 10% over the baseline using our approach. However, in terms of overall dependency parsing and not just with respect to PP-attachment, our system is unable to beat the MSTParser model. However, we need to test our approach on a larger dataset, and across other domains besides Tourism. Besides Hindi, there is also scope for exploring other languages as an aid for pp-attachment disambiguation in English. Our approach could also be used for wh-clause attachment. Since incorrect pp-attachment has a direct consequence on Machine Translation, one interesting analysis could be to use pp-attachments from our system and check for improvement in the quality of translation.

References

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving Parsing and PP Attachment Performance with Sense Information. *ACL*. 317–32. Cite-seer.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Mishra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. *Proceedings of the Third Linguistic Annotation Workshop*. 186-189. Association for Computational Linguistics.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. *Proceedings of the 15th conference on Computational linguistics-Volume 2*. 1198–1204. Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*. 19(2):313-330. MIT Press.

Srinivas Medimi and Pushpak Bhattacharyya. 2007. A Flexible Unsupervised PP-Attachment Method Using Semantic Information. *IJCAI*. 1677–1682.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2009. Multilingual dependency analysis with a two-stage discriminative parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning*. 216-220. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29(1):19-51.

Marian Olteanu and Dan Moldovan. 2005. PP-attachment disambiguation using large context. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 273–280. Association for Computational Linguistics.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. *Proceedings of the workshop on Human Language Technology*. 250–255. Association for Computational Linguistics.

Shiva Reddy and Serge Sharoff. 2011. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*. 11-19. Asian Federation of Natural Language Processing.

Alexander M. Rush and Michael Collins. 2012. A tutorial on dual decomposition and Lagrangian relaxation for inference in natural language processing. *Journal of Artificial Intelligence Research*.

Lee Schwartz, Takako Aikawa, and Chris Quirk. 2003. Disambiguation of English PP attachment using multilingual aligned data. *Proceedings of MT Summit IX* Cite-seer.

- David A. Smith and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. *Proceedings of the Workshop on Statistical Machine Translation*. 23-30. Association for Computational Linguistics.
- Jiri Stetina and Makoto Nagao. 1997. Corpus based PP attachment ambiguity resolution with a semantic dictionary. *Proceedings of the fifth workshop on very large corpora* Citeseer.
- Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. 173–180. Association for Computational Linguistics.